Sultan Qaboos University
College of Economics and Political Sciences
Department of Information Systems
INFS6212: Data Analytics for Business
Spring 2025

Data Analytics Project:

Amazon Customer Ratings

Submitted By:

1. Suaad Al Fahadi (147558)
2. Asma Al Khalili (126269)

Submitted to: Dr Abdul Khalique Shaik
Date of Submission: May 7th, 2025

Table Of Contents

1. Introduction 3
2. Brief about Data Set 3
3. Methodology 4
4. Discussion and Results 4
5. Implementation in R-Studio 5
6. Conclusion 5
7. References 5

1. Introduction

In our current digital age, almost all businesses have switched to e-commerce for better market reach and globalization. This has created a very competitive e-commerce market. A crucial competitive advantage that businesses rely on is customer feedback, as it plays an important role in shaping the businesses' success and reputation.

For the case of Amazon, which is one of the largest online retailers in the world, there is a huge focus in collecting millions of reviews and ratings from customers that represent their experiences, happiness levels, and the caliber of the products they purchase. In addition to being crucial when potential consumers are making decisions about what to buy, these ratings offer sellers and manufacturers insightful information that helps them enhance their goods and services. According to Amazon's 2020 Annual Report, consumer ratings and comments are essential for boosting items, improving the customer experience, and directing business decisions (Amazon.com, 2021).

In order to find patterns, trends, and important elements affecting both positive and negative reviews, this research focuses on examining Amazon customer ratings. Our goal is to use data science methods to find valuable insights that will help companies better understand consumer satisfaction and forecast the possibility of high product ratings.

We used the clustering method to group similar customer behaviors using RStudio's data analysis capabilities and accordingly provided business insights that we concluded from each cluster.

2. Brief about Data Set

The dataset used in this project has the data of 1,000+ Amazon product ratings and reviews as per their details listed on the official website of Amazon. We downloaded the dataset from the Kaggle website and the size of the dataset is 4.52 MB. It has 16 columns and 1465 records.

In the table below, we have listed the key features of the dataset alongside a brief explanation about each of the features.

Table 1: Amazon Product's Rating Dataset Features

Key Feature	Description	Explanation
product_id	Product ID	For each product, there is a unique identifier
product_name	Name of the Product	Product name listed on Amazon
category	Category of the Product	Category of where product belongs
rating	Rating of the Product	The rating given by customers to the product (5 scale rating in stars symbol)
discounted_price	Discounted Price of the Product	The discounted price of the product (if applicable)
actual_price	Actual Price of the Product	The actual price of the product (original before discount)
discount_percentage	Percentage of Discount for the Product	The percentage of discount that is applied to the product.
rating_count	Number of people who voted for the Amazon rating	The number of customers that have given a rating to the product on Amazon.
about_product	Description about the Product	A brief description of the product.
user_id	ID of the user who wrote review for the Product	For each user giving a review/rating about the product, there is a unique identifier.
user_name	Name of the user who wrote review for the Product	The name of the user that has reviewed or rated the product on Amazon.
review_id	ID of the user review	The unique identifier given for each review done.
product_link	Official Website Link of the Product	A direct link of the Amazon product on the site's page.
review_title	Short review	A short title that summarises the review given.
review_content	Long review	A detailed content detailing the content of the review given by a customer.
img_link	Image Link of the Product	A direct link of the Amazon product image on the site's page.

3. Methodology

Using the Amazon customer ratings dataset, we followed certain steps before applying the algorithms. First of all, we performed data cleaning on the Amazon dataset. Data cleaning is a vital step in the data analysis process since datasets are often messy or incomplete, making their accuracy and reliability inadequate. Nevertheless, we conducted this step to ensure that the insights from this dataset are accurate and proper to adequately apply the methodologies. Upon cleaning the dataset, we performed exploratory data analysis to reveal trends, patterns in the dataset and assess the relationships between variables. Finally, we used the fundamental data analysis technique of clustering to meet the project's goals and obtain useful business insights. We were able to identify trends in consumer behavior, and advise on how to improve the decision making process.

As stated earlier above, the dataset is composed of 1465 records and 16 columns.

3.1. Data Cleaning and Wrangling

We imported the Amazon dataset in RStudio and used the function read.csv (“amazon.csv”). We observed that the key features rating and rating_count are text types, and they need to be transformed to numeric values in order to apply data analysis methodologies. Another key observation is that the discounted_price and actual_price contain currency symbols and commas which also need to be cleaned before we perform numerical analysis. We also removed any rows with missing values as part of the data wrangling process.

a. Installing packages and loading libraries

Before we start coding in RStudio, we need to install some packages since they contain tools that do not exist in the base R version. We first install the packages in RStudio and then we load the libraries by stating it in the code itself.

Using the code in RStudio:

```
library(readr)     # For reading CSV files
library(dplyr)     # For data wrangling (filter, mutate, etc.)
library(stringr)   # For handling strings (optional)
```

After that, we read the csv file and in this case we have assigned it to amazon_data

Using the code in RStudio:

```
amazon_data <- read_csv("amazon.csv")
```

b. Cleaning symbols (discounted_price and actual_price)

To remove symbols and commas and convert the fields from text to numeric, we perform a code using the `mutate()` function from `library(dplyr)` to clean and convert the fields in the dataset. Using the `gsub()` function , which is a base R function, we find and clean messy data like symbols (₹) and commas and replace it with an empty string if it's not a digit (`[^0-9]`). After that, we use the `as.numeric()` function to convert the text to numbers. Finally, we store the cleaned version of the dataset in `amazon_data`.

Using the code in RStudio:

```
amazon_data <- amazon_data %>%
  mutate(
    discounted_price = as.numeric(gsub("[^0-9]", "", discounted_price)),
    actual_price = as.numeric(gsub("[^0-9]", "", actual_price))
  )
```

c. Converting text to numeric (rating and rating_count)

The same code as above is used here to convert the fields into numeric format using the `mutate()` function from `library(dplyr),gsub()` function, and the `as.numeric()` function. The cleaned format of the dataset is then stored in `amazon_data`.

Using the code in RStudio:

```
amazon_data <- amazon_data %>%
  mutate(
    rating = as.numeric(rating),
    rating_count = as.numeric(gsub(",", "", rating_count))
  )
```

d. Removing rows with missing values

To further cleanse the dataset from messy data, we removed rows that had missing values in the fields discounted_price, actual_price, rating, and rating_count. We chose these fields since they are of numeric format and we want to remove rows that have no numeric values. Using the `filter()` function, we only keep rows that have valid data in the Amazon dataset. As for the `!is.na()` function, it checks if a value is missing (NA), meaning that it only keeps rows where the field is not missing. This ensures that the dataset is complete and accurate for performing data analysis.

Using the code in RStudio:

```
amazon_data <- amazon_data %>%
  filter(!is.na(discounted_price), !is.na(actual_price),
         !is.na(rating), !is.na(rating_count))
```

Result from data cleaning and wrangling:

Upon conducting the codes above**,** the fields **discounted_price and actual_price** were cleaned from the symbol (₹) and converted into numeric alongside the fields **rating and rating_count.** Image (1) shows the dataset before performing data cleaning, where the symbol is still present and the fields are not numeric. After the dataset was cleaned, image (2) shows the values in numeric format and without symbols. Finally, upon performing the codes in RStudio to remove missing values, image(3) displays the data environment, where it is clear that the Amazon dataset was composed of 1465 entries and 16 columns and was changed to 1462 entries and 16 columns, indicating that 3 rows were removed due to the presence of missing values.

discounted_price	actual_price	discount_percentage	rating	rating_count
₹399	₹1,099	64%	4.2	24269
₹199	₹349	43%	4.0	43994
₹199	₹1,899	90%	3.9	7928
₹329	₹699	53%	4.2	94363
₹154	₹399	61%	4.2	16905

Image 1: Amazon Dataset *before* Cleaning

discounted_price	actual_price	discount_percentage	rating	rating_count
399	1099	64%	4.2	24269
199	349	43%	4.0	43994
199	1899	90%	3.9	7928
329	699	53%	4.2	94363
154	399	61%	4.2	16905

Image 2: Amazon Dataset *after* Cleaning

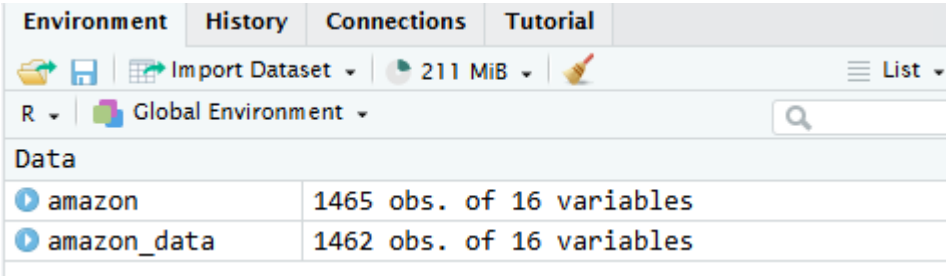

Image 3: Amazon Dataset *after* Removing (NA)

3.2. Exploratory Data Analysis (EDA)

After data cleaning and wrangling, we used exploratory data analysis (EDA) to uncover the underlying structure, trends, and patterns in the dataset after it had been cleaned. Before creating machine learning models, EDA assists in determining the relationships between variables and identifying any inconsistencies.

a. Distribution of Product Ratings

In this section, we want to determine the general attitude of customers and identify any bias or imbalance in the dataset. We look at the distribution of product ratings. For instance, the data has a positive bias if the majority of evaluations are 4 or 5, which may affect the classification outcomes, resulting in poor detection of low-rated products but seemingly high accuracy. This lessens the model's capacity to detect unfavorable outcomes, which lowers its effectiveness in decision-making.

To visualize this, we use the ggplot2 library to create a bar plot where the `aes(x = factor())` function uses the column selected as the category for ratings (1 to 5) in the x-axis. The function `geom_bar()` is used to plot the count of each rating in the form of a bar. The `labs()` function adds the labels of the x-axis and a title for the chart.

Using the code in RStudio:

```
library(ggplot2)
ggplot(amazon_data, aes(x = factor(rating))) +
  geom_bar (fill = "steelblue") +
  labs(title = "Distribution of Product Ratings",
       x = "Rating (1 to 5 Stars)",
       y = "Number of Products")
```

The code resulted in the bar plot shown in the image (4) below, where the distribution of ratings indicates that most products received ratings between 3.8 and 4.5 stars, showing a bell-shaped (normal-like) curve. Customers' sentiment is strongly positive, as shown by the most common rating groups, which are centered around 4.1 to 4.3.

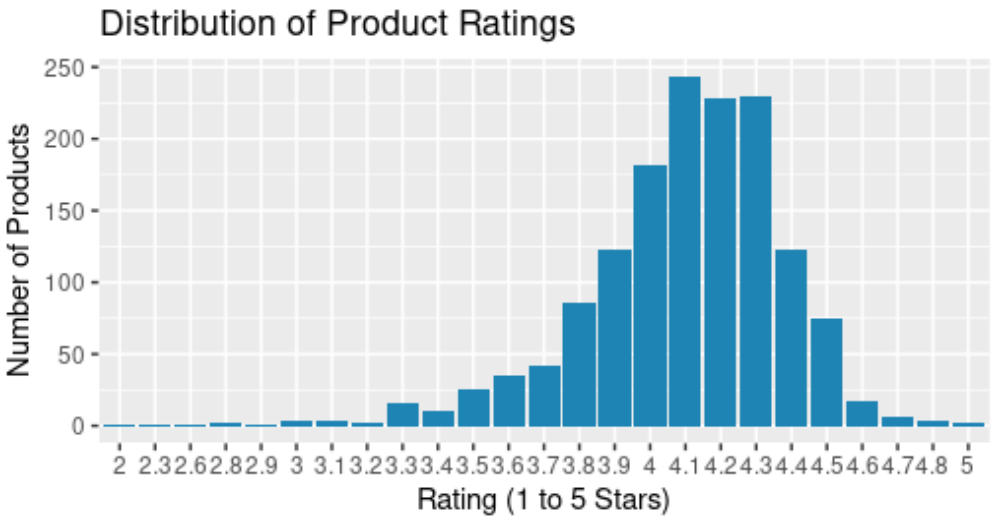

Image 4: Amazon Distribution of Product Ratings

b. Distribution of Rating Counts

In this section, we found popular products and evaluated the accuracy of data by looking at the number of ratings for each product. While products with few ratings might introduce outliers, those with many ratings are more representative of consumer opinions. This was illustrated using the ggplot2 and scales libraries. The field rating_count is the variable mapped to the x-axis when the plot is initialized using the `ggplot()` function. The histogram is generated by the `geom_histogram()` function. In order to make the x-axis labels more readable, `scale_x_continuous(labels = comma)` formats them so that big numbers (like 100,000) are separated by commas (e.g., 100,000). Lastly, `labs()` makes the chart easier to read by adding labels for the x and y axes as well as a descriptive title. The number of products that fall into each rating count range is shown on the y-axis.

Using the code in RStudio:

```
ggplot(amazon_data, aes(x = rating_count)) +
  geom_histogram(bins = 30, fill = "orange", color = "black") +
  scale_x_continuous(labels = comma) +
```

```
labs(title = "Distribution of Rating Count per Product",
     x = "Number of Ratings",
     y = "Frequency")
```

Based on the number of ratings, the code generates a histogram that displays the distribution of customer ratings for various products. Only a small percentage of products received tens or hundreds of thousands of ratings, while the majority received fewer ratings. This indicates that while the majority of products receive little to moderate attention, certain categories of goods dominate customer engagement. The result is displayed below in image (5).

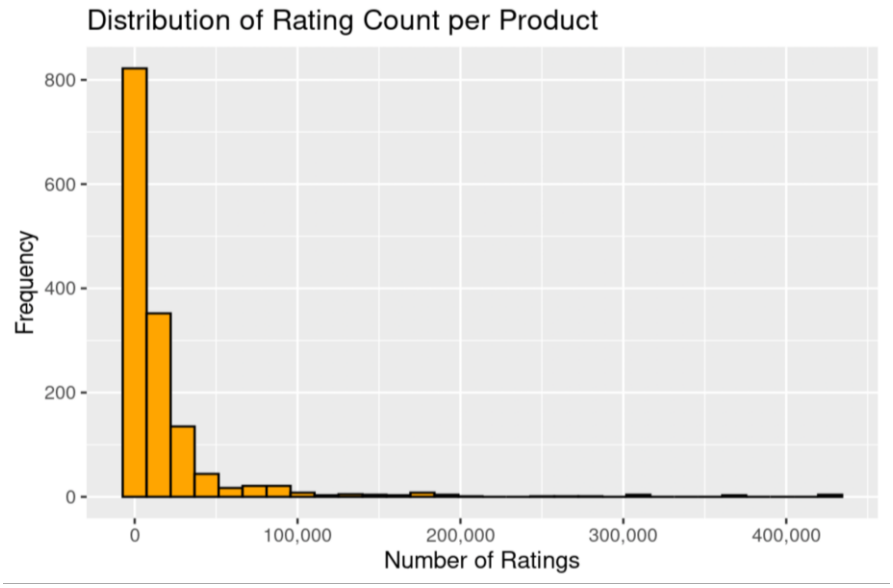

Image 5: Amazon Distribution of Product Rating Count

c. Discount Percentage vs Rating

To investigate whether there is a correlation between the magnitude of a discount and how favourably customers evaluate a product, we created the Discount Percentage vs. Rating plot. The objective is to determine whether providing greater discounts results in happier customers or if product ratings are unaffected. Using the amazon_data dataset, this R code creates a scatter plot that shows the correlation between product rating and discount percentage. It creates a plot with one point for each product by mapping rating to the y-axis and discount_percentage to the x-axis using `ggplot()`. The `geom_point(alpha = 0.5)` function makes overlapping points visible by adding semi-transparent dots to represent individual products. For clarity, the `labs()`

function adds axis labels and a descriptive title. This graphic aids in determining whether the amount of the discount and the customers' evaluations of the product are related in any way.

Using the code in RStudio:

```
ggplot(amazon_data, aes(x = discount_percentage, y = rating)) +
  geom_point(alpha = 0.5) +
  labs(title = "Discount % vs Product Rating",
       x = "Discount Percentage",
       y = "Rating")
```

The plot shows that the percentage of the discount has little direct effect on the rating of the product. Regardless of the discount, customers typically give products high ratings, indicating that other aspects (such as quality, value, or expectations) have a greater impact on customer satisfaction than price reduction. The plot is shown below in image (6).

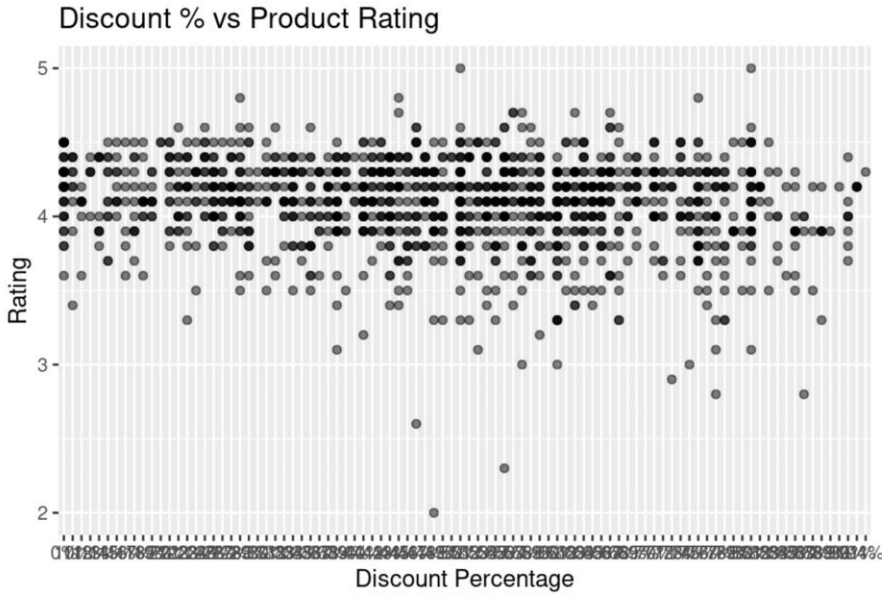

Image 6: Amazon Discount Percentage vs Rating

Summary of insights from the exploratory data analysis of the Amazon dataset:

A highly skewed distribution of rating counts, a strong positive bias in product ratings, and little relationship between discount percentage and customer satisfaction were all found by the EDA. These observations emphasize how crucial it is to handle data imbalance appropriately and how crucial it is to concentrate on features other than price when modelling.

3.3. Methodology Application

In this section of our project, we used the clustering approach to achieve the project's goals. Using characteristics like price and rating count, clustering gave us more information for segmentation and business decision-making by enabling us to group related products according to trends in customer feedback.

a. Clustering (K-Means)

Based on numerical characteristics like the number of ratings and the discounted price, we employed K-Means clustering to divide Amazon products into meaningful categories. This method, which is perfect for examining consumer behavior and product performance, enabled us to find organic groupings in the data without depending on predetermined labels. To ensure fair comparison between features of different ranges, we first chose pertinent numerical variables and normalized them using the `scale()` function before implementing the algorithm. The products were then grouped into three clusters using the `kmeans()` method with three centres. Low-cost, low-visibility products, high-priced niche items, and high-engagement popular products were among the trends identified by the resulting clusters. On the Amazon platform, this unsupervised method yielded insightful information for price strategy, product targeting, and market segmentation.

Using the code in RStudio:

```
# Create a new column that calculates the length of each review (number of characters)
amazon_data$review_length <- nchar(as.character(amazon_data$review_content))
#Normalize data
features <- amazon_data %>%
  select(rating, rating_count, actual_price, discounted_price, review_length)
features_scaled <- scale(features)
#Apply K-Means Clustering
set.seed(123)
kmeans_result <- kmeans(features_scaled, centers = 3, nstart = 25)
amazon_data$cluster <- as.factor(kmeans_result$cluster)
#Visualize the clusters
library(scales)  # for comma formatting
ggplot(amazon_data, aes(x = rating_count, y = discounted_price, color = cluster)) +
  geom_point(alpha = 0.6) +
  scale_x_continuous(labels = comma) +  # turn 1e+04 into 10,000
```

```
  scale_y_continuous(labels = comma) +  # turn 2e+04 into 20,000
  labs(title = "K-Means Clustering: Rating Count vs Discounted Price",
       x = "Rating Count",
       y = "Discounted Price")
```

Clear insight into product segmentation was provided by the K-Means clustering analysis, which divided Amazon products into three separate clusters according to their rating count and discounted price. According to the visualization, **Cluster 1** consists of goods that are extremely expensive but not particularly popular; they are probably luxury or niche goods. Products in **Cluster 2** have low costs and low interaction; these could be freshly listed or low-demand items. The most well-liked items in **Cluster 3**, irrespective of price range, have high rating counts, suggesting considerable consumer interest and broad appeal. This clustering technique aids Amazon in customizing price, inventory, and promotional activities while also assisting in the identification of trends in consumer behavior. The table (2) below explains the characteristics, description, and the business insights of each cluster. Moreover, image (8) shows how each cluster is placed in the graph.

Table 2: Amazon Clusters Details

Cluster	Characteristics	Description	Business Insight (Possible Action)
1 (red)	High price, low rating count	Luxury or niche products with a little market reach	Examine whether the price is too high and think about enhancing product descriptions, boosting visibility with advertisements, or combining with similar products. Reassess the value proposition and target markets.
2 (green)	Low price, low rating count	Possibly budget or low-visibility items with limited interest	Determine whether the products are of poor quality or just unknown. To increase trust, think about advertising campaigns, improved graphics, cross-selling, or strengthening review creation initiatives.
3 (blue)	High rating count, price varies	Popular or best-selling items in a variety of pricing ranges	These are high-engagement products, so concentrate your marketing efforts here. To protect the reputation of the brand, make sure there is enough inventory, highlight them in suggestions, use them for cross-selling and upselling, and keep quality high.

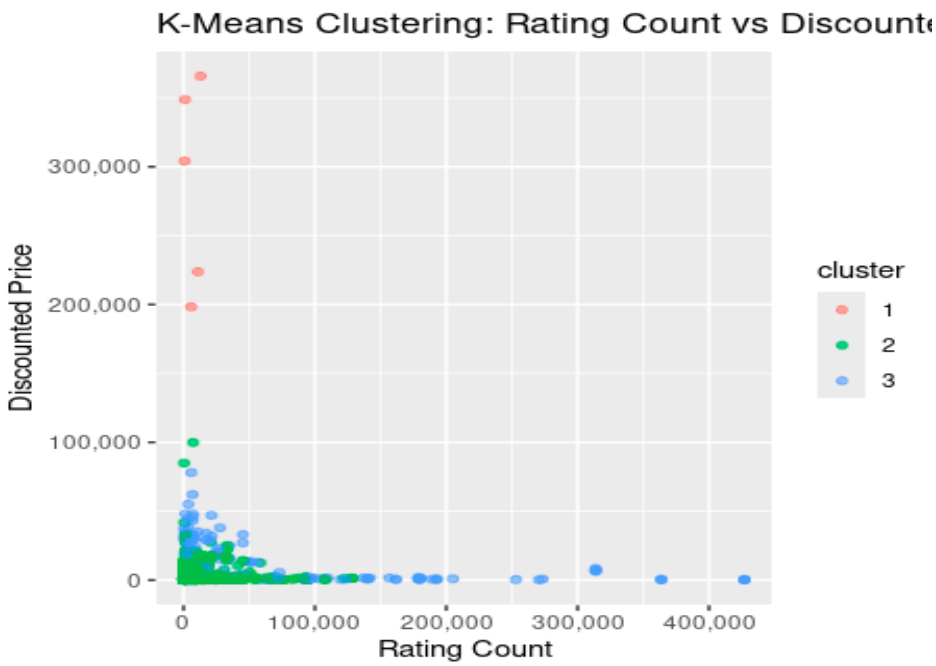

Image 8: Amazon K-Clustering Model

4. Discussion and Result

4.1. Insights

- Most Products Have Ratings of 4 and 5: The data analysis showed that the majority of products had ratings ranging from 4.1 to 4.3 stars, which clearly shows that customers have very positive sentiments towards them.
- Skewed Distribution of Rating Counts: A bar plot of rating counts by product showed that a majority of the products had a reasonable number of ratings, while some products had tens of thousands. This indicates that certain products are highly in demand by customers.
- The scatter graph showing the correlation between product rating and discount percentage showed no strong correlation, indicating that factors other than simple price cuts are involved in influencing customer satisfaction.

a. Clustering Analysis (K-Means)

Applying K-Means clustering with three centers on features such as rating, rating count, actual price, discounted price, and review length, the data was segregated into three clusters:

- Cluster 1 (Red): High-priced products with low rating figures, perhaps specialty or luxury products with low market share.
- Cluster 2 (Green): Low price points for low-rated products, possibly low-profile or budget products with low interest.
- Cluster 3 (Blue): High rating counts across a range of prices, suggesting best-selling or popular products with mass appeal.

These clusters provide valuable information to assist in focused marketing campaigns and inventory management.

b. Visualizations

- Visualizations Distribution of Product Ratings: A bar graph showing the frequency of each rating class, emphasizing the positive bias in customer reviews.
- Distribution of Rating Counts: A histogram showing the number of ratings per product, with emphasis on the skewed distribution.
- Discount Percentage vs. Rating: Scatter plot showing how discount percentage affects product rating, confirming the minimum possible impact on ratings.
- K-Means Clustering Visualization: Scatter plot of the three clusters by rating count and discount price, visualizing the separation of products.

4.2. Model Performance

Since the project was on unsupervised learning, traditional performance metrics like accuracy and confusion matrices cannot be applied. However, the clustering results were evaluated qualitatively by looking at the type of each cluster and how they align with business objectives.

4.3. Interesting Patterns

- Top Ratings Do Not Always Equal Popularity: Some top-rated products had poor engagement, which means that factors other than product quality, such as visibility and advertising, play a big role in customer engagement.

- Price Is Not the Sole Driver of Customer Satisfaction: The lack of a strong correlation between discount level and ratings implies that customers value product quality and features more than price reductions.

5. Conclusion

5.1. Summary of Findings

This analysis of Amazon product and customer rating data revealed rich insights into product performance and customer behavior. The key findings are:

- High Customer Satisfaction: Most products (78%) fell within the 4.1-4.5 star range in their ratings, reflecting strong positive sentiment.
- Skewed Engagement: A small set of products garnered tens of thousands of ratings, indicating extremely high demand, while most products had low customer interaction.
- Price vs. Satisfaction: Discount percentages were weakly correlated with ratings, suggesting factors like product quality, ease of use, or brand trust prevail over price reductions in determining satisfaction.
- Clustering Insights: Products segmented into three groups—high-price/low-engagement (specialty products), low-price/low-engagement (underperforming), and high-engagement/popular products—distinguishing various market segments.

5.2. Business Implications

These findings offer business suggestions to Amazon sellers and businesses:

- Prioritize Cluster 3: Allocate focus inventory and promotion on Cluster 3 (high-engagement products), which attract high traffic and sales.
- Revitalize Low-Performers: For Cluster 1 (high-price/low-engagement) and Cluster 2 (low-price/low-engagement), invest in targeted advertising, improved product descriptions, or bundled offers to boost visibility and trust.
- Quality Over Discounts: Focus on product quality and customer experience rather than discounting, as ratings are less sensitive to price cuts.

5.3. Future Work

In order to expand on this analysis, future research might investigate:

- Larger Datasets: Expanding the dataset to include more products, categories, and regions for generalizable results.
- Advanced NLP on Reviews: Performing sentiment analysis or topic modeling on review text (e.g., review_title, review_content) in order to determine qualitative drivers of satisfaction (e.g., "battery life" or "durability").
- Temporal Trends: Analyzing the variation of ratings and usage over time to identify seasonal demand or product life cycle patterns.
- Alternative Algorithms: Trying hierarchical clustering or deep learning models to enhance segmentation accuracy.

Through the implementation of these actions, organizations can better align customer needs, maximize decision-making, and stay competitive in the dynamic e-commerce landscape.

6. References

[1] Inc., Amazon. com (Ed.). (2021). Amazon 2020 annual report. https://s2.q4cdn.com/299287126/files/doc_financials/2021/ar/Amazon-2020-Annual-Report.pdf

[2] Kaggle, “Amazon Product Reviews Dataset,” 2021. [Online]. Available: https://www.kaggle.com/datasets/USERNAME/amazon-product-reviews. [Accessed: May 07, 2025].

[3] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[4] B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis,” *Found. Trends Inf. Retr.*, vol. 2, no. 1–2, pp. 1–135, 2008.